Using unmarked contexts in nominal lexical semantic classification

Lauren Romeo[*], Sara Mendes[*,†], Núria Bel[*]
*Universitat Pompeu Fabra, Roc Boronat, 138, Barcelona, Spain
†Centro de Linguística da Universidade de Lisboa, Av. Prof. Gama Pinto, 2, Lisboa, Portugal
{lauren.romeo,sara.mendes,nuria.bel}@upf.edu

Abstract

The work presented here addresses the use of unmarked contexts in pattern-based nominal lexical semantic classification. We define unmarked contexts to be the counterposition of the class-indicatory, or marked, contexts. Its aim is to evaluate how unmarked contexts can be used to improve the accuracy and reliability of lexical semantic classifiers. Results demonstrate that the combined use of both types of distributional information (marked and unmarked) is crucial to improve classification. This result was replicated using two different corpora, demonstrating the robustness of the method proposed.

1 Introduction

Lexical resources annotated with lexical semantic classes have been successfully incorporated into a wide range of NLP applications, such as grammar induction (Agirre *et al.*, 2011) and the building and extending of semantic ontologies (Abbès *et al.*, 2011). However, lexical semantic tagging in large lexica is mostly done by hand, implying high costs with regard to maintenance and domain tuning. As the use of an inadequate lexicon is one of the causes of poor performance of NLP applications, current research to improve the automatic production of rich language resources, and of class-annotated lexica, in particular, is critical.

One way to approach this task is through supervised cue-based lexical semantic classification. Based on the distributional hypothesis (Harris, 1954), according to which words occurring in the same contexts can be said to belong to the same class, cue-based lexical semantic classification uses particular linguistic contexts where nouns occur as cues that represent distinctive distributional traits of a lexical class. Yet, training a classifier with information about word occurrences in a corpus within a selected number of contexts can present a challenge, mainly because specific words might be observed in a number of class-indicative contexts but not always are.

This type of marked, or class-indicative, context (e.g. co-occurrence with specific prepositions, predicate selectional restrictions, and grammatical information, such as indirect objects) are sparse in any corpus as, being so specific, they do not occur often with each target noun. Using only exclusive class-indicative contexts as features in nominal lexical semantic classification has been shown to not always provide sufficient information to make a decision regarding class membership of a noun (Bel *et al.*, 2012), especially when the data does not contain relevant co-occurrences or when those co-occurrences are too disperse to be correlated.

Recent work on the use of distributional models for nominal classification tasks (Romeo *et al.*, 2014) discusses potential bottlenecks of models using data extracted with lexico-syntactic patterns as features, identifying data sparsity as one of the major issues affecting the performance of these systems. In fact, the selection of class-indicative information, in an attempt to provide relevant information to classifiers and thus reduce noise, naturally limits the amount of data available to the system, often resulting in sparse vectors.

Resulting from the necessity of selecting the information provided to classifiers, in an attempt to improve the accuracy of classification decisions, the sparse data problem in nominal lexical semantic classification is one of the crucial issues to be addressed to improve the performance of these systems. We propose to approach this issue by utilizing a larger fraction of the distributional information available in a corpus, by incorporating information typically considered non-indicatory of semantic class membership, which we will designate as *unmarked contexts* (see Section 3 for a definition).

Our hypothesis is that the distributional behavior of nouns of a particular class in this type of generally occurring contexts can show certain characteristics in common that may be explored in the context of lexical semantic classification. Our goal in the work presented here is to evaluate to which extent this information, in combination with the widely explored class-indicative lexico-syntactic con-

texts, can be used to improve results in classification tasks, by providing more information to classifiers. To do this, we experiment with English nouns of the following lexical semantic classes: INFORMATION (INF), ORGANIZATION (ORG), LOCATION (LOC), EVENT (EVT) and HUMAN (HUM).

The work presented in this paper is structured as follows: Section 2 describes theoretical claims used in approaches to nominal lexical semantic classification, as well as related work; Section 3 elaborates upon the concept of unmarked contexts; Section 4 describes the methodology followed; Section 5 presents the results obtained; Section 6 discusses their implications; and Section 7 concludes with final remarks and future work.

2 Motivation and Related Work

In semantic classification approaches grounded in usage-based theories of grammar (Goldberg, 2006), a lexical class is seen as a generalization of the systematic co-distribution of a number of words and contexts. Construction-based grammar hypotheses allow us to predict there are sets of word occurrences that, together, constitute a class mark, indicating a particular semantic class, in line with the structuralist notion of markedness (Jakobson, 1971).

The identification of relevant cues for machine learning classification is problematic as low frequency evidence is typically disregarded by automatic systems. To overcome this problem, Bel (2010) applied smoothing methods, demonstrating increases in accuracy, though low frequency words remained problematic for classifiers when evidence was scarce, and thus not considered as a positive cue for the class, although it was indicative.

In Bel *et al*. (2012) we built on this hypothesis, assuming only frequently occurring contexts would be efficient for classification tasks, and considering only frequent predicates, prepositions, affixes, etc., as well as negative cases (i.e. marked cues for other classes), as indicators for a class membership decision. Thus, we ignored all information that, although frequent, co-occurred with nouns from all classes and was deemed not distinctive of a particular class, as well as all information that, though distinctive, was not frequent enough to be used by the classifier.

Table 1 provides examples of contexts considered in that work, which did not rely on unique, exclusive hints, but a number of them that, when correlated, could identify members of a given class. However, our results were inconclusive, a fact which we attributed to the high impact of sparse data.

Class	Examples of lexico-syntactic patterns
ORG	x-NN (found|establish|organize)-VBD
LOC	(inside|outside)-IN (the|a|an)-(DT|Z) x-NN
INF	(submit|publish|report)-V* (the|a|an)-(DT|Z) x-NN
EVT	during-IN (the|a|an)-(DT|Z) x-NN
HUM	x-(-er|–or| –man)-NN

Table 1: Examples of lexico-syntactic patterns indicative of 5 different lexico-semantic classes, which we refer to as *marked contexts*.

According to Bybee (2010), general contexts, not exclusive to a particular class (i.e. unmarked contexts, as defined in Section 3), are more frequent than contexts marked toward a particular class, as they occur with nouns of all classes. In view of this, it becomes apparent that a large part of available distributional data is not taken into consideration when these very general co-occurrences observed with nouns of all classes (e.g. co-occurrence with an article) are treated as stop words or contexts in lexical semantic classification tasks.

The basic claim leading most authors to neglect this kind of context is that, due to its assumed undifferentiated distribution, this information presents a challenge for classifiers to accurately use it in class membership decisions, which is bound to negatively affect results (see Cooke and Gillam (2008), Turney and Pantel (2010), Bullinaria and Levy (2012), among many others). In contrast with this mainstream position, Rumshisky *et al*. (2007) argued there is an asymmetry in the way certain word senses are used in language, preferably or rarely occurring in certain very general contexts (e.g. subject position, occurrence with an adjectival modifier, etc.).

This type of asymmetry is essentially referring to a difference in how general, semantically neutral distributional contexts are more or less frequent in data, depending on the sense in which a word is

used. We hypothesize that these tendencies can also be observed when considering different lexical semantic classes.

In this paper, we propose a way to include this type of asymmetry in the information provided to classifiers to verify its impact on their overall performance. Considering such distributional evidence will increase the amount of information made available to classifiers. Our main claim is that devising a strategy to informatively include this type of distributional information in classification tasks can allow us to take advantage of a bigger portion of the data available in corpora and improve the accuracy of classifiers in this way.

3 Unmarked contexts

In contrast with mainstream approaches to cue-based lexical-semantic classification, we argue for the inclusion of a type of distributional information typically not considered to be indicatory of class membership, and thus not informative to automatic classification systems. These are very general contexts of occurrence typically disregarded as semantically-empty and thought to be too general to contribute any relevant information. At the same time, they correspond to a large amount of corpus data that is a priori not considered due to the assumption that it does not provide any information.

Examples of such distributional information regard whether nouns occur preceeded by an article, in singular or plural form, whether they are a head or a complement in NPs containing the preposition *of*, if they occur as the subject of the verb to *be*, etc. We will henceforth designate these contexts as ***unmarked contexts***, the counterposition of the class-indicatory contexts, i.e. ***marked contexts***, used as class marks in cue-based lexical-semantic classification systems.

Following the conclusions of Rumshisky *et al.* (2007) regarding asymmetries in the distribution of word senses in general contexts, our hypothesis is that the distribution of members of a class with respect to their occurrence in particular unmarked contexts is consistent and thus can be captured and used to inform classifiers and improve results, when considered along with other indicatory, or marked, contexts.

We also hypothesize that unmarked contexts will alleviate problems caused by data sparsity in classification tasks by providing additional information to classifiers. To assess to which extent this information can be used in classification tasks, we had to identify such contexts and verify whether our hypothesis was confirmed, i.e. if they showed significant variations in terms of distribution that might be explored to augment the amount of information made available to classifiers[1]. Additionally, and given the specific properties of this type of distributional information, we also had to define a strategy to informatively provide it to classifiers (see Section 3.2 for details).

3.1 Identifying unmarked contexts

Considering the characteristics of the contexts discussed above, we identified 32 unmarked contexts under a frequency criterion (see Table 2 for a description of the different contexts identified), hypothesizing that more frequent contexts combine with more nouns in the corpus and thus should not be marked for any restricted set. However, although they are not considered to be class marks, we expect these contexts to be asymmetrically distributed between lexical semantic classes, in an analogous way to what was observed by Rumshisky *et al*. (2007) with regard to the distributional behavior of different word senses in language use.

We studied the distribution of these contexts in a web-crawled corpus (see Section 4), comparing the distribution of each context over all the nouns in the corpus and over nouns defined as part of a specific lexical semantic class, according to a gold standard (see Section 4.2). To do this, we calculated the mean of occurrence of nouns pertaining to a particular class in a specific unmarked context, as well as the mean of occurrence of all the nouns in the corpus in that same context; we then determined

[1] The approach detailed in this paper contrasts with a 'bag of words' approach to classification as, even in the case of what we call unmarked contexts, we rely on cue information to populate our vectors. Thus, the information provided to classifiers takes into consideration linguistic information, such as syntactic order or dependencies. Moreover, our use of linguistically-motivated features, from the inherently distinctive to the more generic, reduces the amount of data needed to obtain a desired level of performance.

if there was a statistically significant difference[2] between the behavior of nouns from specific classes and the behavior of nouns in general with regard to the contexts identified as unmarked.

Feature Type	Description	Examples
article	target noun preceded by a(n) (in)definite article	*(a|an)-(DT|Z) x-NN* or *(the)-(DT|Z) x-NN*
number	target noun in plural/singular form	*x-NNS* or *x-NN*
copula	target noun as subject/object of verb *to be*	*x-NN* be-*VBZ* or be-*VBZ x-NN*
modifiers	adjective or nominal modifier preceding target noun	*x-JJ x-NN* or *x-NN x-NN*
preposition *of*	target noun preceding/following the preposition *of*	*x-NN of-IN* or *of-IN x-NN*
subject of V	target noun as subject of each of the 20 most frequent verbs in the corpus	*x-NN (have|get|make|see|do|take|go|use|find|help|read|know|provide|give|keep|come|say|create|visit) -VB(Z|D)*

Table 2: Description of unmarked contexts identified and used in our experiments.

The results showed there were statistically significant differences ($p<0.05$) in the behavior of nouns in particular classes with regard to certain unmarked contexts. For instance, the occurrence of INF, ORG, LOC, and HUM nouns with a definite article (*the-DT*) showed to be significantly different from its average occurrence with all the nouns in the corpus. The occurrence with an indefinite article (*a|an-DT*), on the other hand, showed to be significantly different for LOC nouns, while the co-occurrence with an adjective (*x-JJ*) was significantly different for INF nouns.

3.2 A strategy to encode unmarked context information in feature vectors

The preliminary study mentioned in Section 3.1 provided evidence confirming that there are, in fact, differences in the behavior of particular lexical semantic classes with regard to their occurrence in unmarked contexts. Thus, the next step consisted in determining the best way to make this information available to classifiers.

Aiming to check the validity of our hypothesis in general, the results obtained in the aforementioned study were not used directly to narrow down the information to include in the vectors used by the classifiers to avoid the risk of over-fitting. Moreover, what was at stake, considering our theoretical hypothesis, was to devise a strategy to account for specific differences between the behavior of each noun considered for classification and the average behavior of all nouns in the corpus with regard to each context considered. Thus, information regarding all 32 unmarked contexts was provided to the classifiers for all lexical classes considered.

To mirror the specificity of the distribution of each noun with regard to each context considered, we subtracted the mean of occurrence of nouns in each context from the actual occurrences of the target noun represented by the vector in that same context to obtain each feature *f*, as defined in Equation 1, where c_i represents a given context, *t* a target noun, *n* any noun belonging to *N*, the set of all nouns in the corpus, and *freq* frequency of occurrence (e.g. $freq(t \mid c_i)$ = frequency of occurrence of the target noun *t* in context c_i).

Equation 1: $$f = \frac{freq(t \mid ci)}{freq(t)} - \frac{1}{|N|} \sum_{n \in N} \left[\frac{freq(n \mid ci)}{freq(n)} \right]$$

Using the difference between the number of occurrences of a given noun and the average occurrence of all nouns in a specific context, we encode the deviation of the behavior of that noun with regard to the general behavior of all nouns in the corpus, under the hypothesis that nouns of the same

[2] In this work, statistical significance was calculated using Student's t-test (cf. Krenn and Samuelsson, 1997).

class display similar tendencies in terms of deviant behavior in the contexts considered, providing relevant information to the classifier. We apply our method to two different corpora making apparent its robustness.

4 Experimental design and setup

In order to evaluate the impact of using distributional information on unmarked contexts for lexical-semantic classification tasks, first, we had to extract distributional information regarding the unmarked contexts identified (see Section 3.1), as well as distributional information regarding class-indicative marked contexts. In our experiments, we used the marked contexts identified and described in Bel *et al*. (2012) (see Table 1 for examples).

Once the distributional information was extracted, we incorporated it in feature vectors, using the different aforementioned strategies for encoding distributional information regarding marked and unmarked contexts, respectively, as detailed further below in this section. Once all of the information was compiled, the vectors were provided to classifiers.

As previously mentioned, our experiments covered English nouns of the classes: INF, ORG, HUM, EVT and LOC (see Section 4.2). For the purpose of the work presented here, we experimented with two corpora to determine the transferability and robustness of our method, independently of specific corpus data.

We first used a general web-crawled corpus (Pecina *et al.*, 2011) consisting of 30 million PoS-tagged English tokens (henceforth Corpus A) to identify unmarked contexts (see Section 3.1) as well as to train our classifiers.

We also employed an excerpt of the web-crawled UkWaC corpus (Baroni *et al.*, 2009), consisting of 60 million PoS-tagged English tokens (henceforth Corpus B) to test our approach on unknown data, in this way ensuring that our approach and classifiers are not over-fitted to any specific corpus, instead confirming that the method we propose can be generalized, and the results obtained are replicable given any dataset.

Regular expressions over both corpora were used to identify occurrences of nouns in marked and unmarked contexts. For marked contexts, the relative frequency of each pattern seen with a particular noun was stored in an *n*-dimensional vector.[3] The occurrences of a noun in unmarked contexts were encoded in the same vectors following the strategy outlined in Section 3.2 (see Equation 1).

4.1 Classification

For classification, we used the Logistic Model Trees (LMT) (Landwehr *et al.*, 2005) Decision Tree (DT) classifier in the WEKA (Witten and Frank, 2005) implementation in a 10-fold cross-validation setting. We conducted binary classifications, one for each semantic class considered. We measure the success of our approach in regards to the joint performance of individual classifiers in accurately distinguishing members of each individual class from any other noun. This method was used in the classification experiments over both corpora described above.

4.2 Gold Standard Description

In regards to the gold standard lists used for training and evaluation, we automatically extracted from WordNet (Miller *et al.*, 1990) all of the nouns encoded in this repository of lexical information that contained a sense corresponding to a class considered in our experiments (e.g. *people* in the case of HUM).

The gold standards were not contrasted with the actual occurrences of the nouns in the corpora. They were, however, balanced with respect to class members and elements not belonging to the class, resulting in the dataset described in Table 3. Each noun appears *x* times in any corpus considered. The elements not belonging to a class were randomly selected from the set of nouns that do not contain a sense in WordNet that corresponded to the target class being classified.

For a fair comparison, the baseline classification model was obtained using the context patterns described in Bel *et al*. (2012) with the LMT classifier, using the previously described gold standard lists over Corpus A. This baseline allows us to assess the impact of unmarked contexts in nominal lexical semantic classification, since the classifiers proposed here that are provided with information on the

[3] In this work, *n* is equal to the amount of marked contexts plus unmarked contexts considered for each class.

distributional behavior of nouns in unmarked contexts also use Bel *et al*. (2012)'s context patterns to extract class-indicative, or marked, distributional information regarding the nouns to classify.

Class	**ORG**	**LOC**	**EVT**	**INF**	**HUM**
Class members	138	157	260	262	246
Elements not belonging to the class	135	156	260	259	246

Table 3: Number of nouns included in gold standards per class.

5 Results

Tables 4 and 5 show results obtained in our experiments in terms of Precision (P), Recall (R) and F-Measure (F). The overall accuracy of all classifiers for each experiment is also provided. The baseline classifiers achieve an average accuracy of 70.84%. By including unmarked contexts in the vectors provided to the classifiers, the average accuracy of the classifiers rises to 75.16%, representing an error reduction of 4.32 points. We tested the statistical significance ($p<0.1$) of this increase in the accuracy of classification and, for all classes except for HUM, the increase in accuracy between the baseline results and those obtained when including unmarked contexts is significant. These results are discussed in detail in Section 6.

Knowing a potential downside of using unmarked contexts in classification tasks is an increase in noise (see Section 6.1 for a detailed discussion regarding this concept), we conducted an error analysis of the results obtained, which made apparent that most of the noise was due to imprecise information extracted with our regular expressions, leading us to revise them. As these revisions resulted from the observation that a portion of the errors in the baseline results was due to imprecise regular expressions, they did not consist in the definition of new marked contexts, rather in a revision of how to extract marked contexts already considered in this work from corpora data. Thus, these revisions resulted in more accurate and better defined regular expressions.

As indicated by the results, these revisions in combination with the unmarked contexts further raised the average accuracy of the classifiers to 76.35% (see Table 4), representing an error reduction of 5.51 points with regard to the baseline. Having obtained these promising results over the data in the corpus used to develop our approach (Corpus A), it was crucial to verify the replicability of our method using a different and completely independent corpus, as described in Section 4. Moreover, replicating the original experiments over a different corpus was also crucial to assure that the revisions made to the regular expressions did not result in any over-fitting between the extraction of distributional information and the corpus being used. The results obtained for the experiments conducted over Corpus B are presented in Table 5.

Class	**baseline**			**baseline + unmarked contexts**			**marked contexts**			**marked + unmarked contexts**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
ORG	0.64	0.62	0.60	0.70	0.68	0.68	0.76	0.74	0.74	0.75	0.74	0.74
LOC	0.72	0.70	0.70	0.73	0.73	0.73	0.70	0.70	0.70	0.77	0.79	0.77
EVT	0.70	0.68	0.67	0.74	0.73	0.72	0.73	0.72	0.64	0.73	0.72	0.69
INF	0.67	0.66	0.65	0.74	0.73	0.73	0.71	0.70	0.69	0.71	0.71	0.71
HUM	0.86	0.84	0.86	0.87	0.86	0.86	0.87	0.87	0.87	0.85	0.84	0.84
Acc	70.84%			75.16%			75.05%			76.35%		

Table 4: Precision (P), Recall (R), and F-Measure (F) of classifiers over Corpus A.

The classifiers that include unmarked contexts yielded an average accuracy of 76.03% over Corpus B, representing an error reduction of 3.34 points with regard to the classifier including only marked contexts (using the revised version of Bel *et al*. (2012)'s cues), which is a statistically significant improvement ($p<0.05$). Moreover, these results represent an improvement of accuracy by 5.19 points with regard to the baseline. This demonstrates, on the one hand, that the definition of relevant contexts based on Corpus A data did not result in an over-fitted approach; and, on the other hand, that the method presented here is robust, as we used our classifiers over a completely different corpus (cf. Sec-

tion 3) and still yielded comparable results. Due to space limitations, below we detail only the results obtained on Corpus B data, as these are independent of all the preliminary studies conducted and thus demonstrate the potential applicability of our approach to any corpus.

Class	marked contexts			marked + unmarked contexts		
	P	R	F	P	R	F
ORG	0.72	0.69	0.69	0.76	0.76	0.76
LOC	0.74	0.71	0.71	0.75	0.75	0.75
EVT	0.68	0.67	0.67	0.73	0.73	0.73
INF	0.69	0.69	0.68	0.70	0.70	0.70
HUM	0.86	0.86	0.86	0.84	0.84	0.84
Acc	72.69%			76.03%		

Table 5: Precision (P), Recall (R), and F-Measure (F) of classifiers over Corpus B.

Class	marked contexts				marked + unmarked context			
	members		non-members		members		non-members	
	P	R	P	R	P	R	P	R
ORG	0.79	0.52	0.65	0.86	0.78	0.72	0.75	0.80
LOC	0.82	0.55	0.66	0.73	0.78	0.70	0.73	0.80
EVT	0.73	0.57	0.63	0.78	0.74	0.72	0.72	0.73
INF	0.72	0.62	0.66	0.75	0.72	0.65	0.68	0.74
HUM	0.87	0.84	0.84	0.87	0.86	0.82	0.82	0.86

Table 6: Precision (P) and Recall (R) of classification of members and non-members of different lexical classes over Corpus B

Table 6 presents the precision and the recall of each individual classifier over Corpus B both with regard to the members of a given class, and those nouns that are not members of that class. This table allows us to identify more precisely how the unmarked contexts contribute to the error reduction in classification.

According to the results, unmarked contexts allow us to gain an average of 10.2 points in recall for class members, demonstrating that they provide useful information to classifiers, which allows them to cover cases which they were not able to before, most likely due to phenomena such as data sparsity. However, the impact on precision varies between classes, as the inclusion of very frequent information in the vectors representing target nouns may provide additional noise to the classifier (see Section 6.1).

The precision of classification of class members decreases slightly with the inclusion of unmarked contexts, although the differences are not statistically significant ($p<0.1$). However, the precision of the classification of nouns not belonging to the classes considered significantly increases ($p<0.1$) with the inclusion of unmarked contexts in all cases except for HUM. This shows that although unmarked contexts do not contribute to a better definition of the characteristics of individual classes (see Table 6), they allow for a cleaner discrimination of members and non-members of a class, contributing to a better partition of the classification space.

Class	marked contexts		marked + unmarked contexts	
	FN (%)	FP (%)	FN (%)	FP (%)
ORG	23.32	6.71	13.43	9.98
LOC	22.30	5.75	14.74	9.71
EVT	21.91	10.42	13.82	12.97
INF	18.94	12.00	17.26	12.63
HUM	7.79	6.01	8.90	6.45

Table 7: Percentage of False Negatives (FN) and False Positives (FP) in classifiers over Corpus B with and without unmarked contexts.

Table 7 presents the percentage of False Positives (FP), i.e. nouns incorrectly marked as members of the class, and False Negatives (FN), i.e. nouns incorrectly marked as not belonging to a class, in the results of each classifier both with and without the inclusion of unmarked contexts. Again, for each of the classes, except HUM, the inclusion of unmarked contexts decreases the percentage of FN, mirroring a reduction in silence. Yet, there was an increase of FP across all classes, signifying an increase of the noise provided to the classifier. These results are discussed in detail in Section 6.

6 Discussion

In Section 5, we presented the results obtained in our experiments using distributional information regarding both marked and unmarked contexts for the classification of English nouns. Overall, our results show that unmarked contexts either improve accuracy or do not affect classification results. Specifically, the improvements in accuracy are particularly significant for those classes for which there were difficulties to find enough occurrences in marked contexts in previous experiments, i.e. those classes with a higher level of FN when classified without using unmarked contexts. This way, the results confirm our general hypothesis that the distribution of words in unmarked contexts, when considered along with contexts marked towards a lexical semantic class, provides information to improve classifiers, particularly when not enough class-specific information is available. In this section we analyze the results obtained, making apparent the main advantages of our proposal.

6.1 A trade-off between silence and noise

An important result of our experiments is the overall reduction in the negative effect of silence in our classifiers, which decreased by an average of 5.21% (see the difference in terms of FN in Table 7), resulting in an increase in accuracy (see Table 5): as more information is supplied to the classifier, the additional information permits more accurate membership decisions. To illustrate this, we consider examples from the INF, ORG and EVT classes, for which there was not enough information for classification when unmarked contexts were not considered. The inclusion of unmarked contexts provided information resulting in correct classifications.

The INF noun *theorem* illustrates this case: *theorem* occurs 118 times in the corpus, though only 8 times in marked contexts, which was not enough to accurately classify it as a member of the INF class. As this noun occurs in class-marked contexts, but not enough times for the classifier to make an accurate prediction regarding its class membership, we can consider that the lack of enough information provided to the classifier is responsible for its misclassification. However, after the inclusion of information regarding the behavior of this noun in unmarked contexts, the classifier was able to accurately decide for its inclusion as a member of the INF class. This was also observed in the case of the ORG noun *secretariat* and the EVT noun *impulse*, which occur 190 and 154 times, respectively, in the corpus, yet only 8 and 12 times in marked contexts, which was not enough for an accurate classification. Again, the inclusion of information regarding the distribution of these nouns in unmarked contexts provided the classifier with sufficient information to allow for correct classification.

One of the main concerns regarding the use of unmarked distributional information was the introduction of extra noise as a side effect and the way this affects classification results. For the purpose of the work presented in this paper, we define noise as contradictory distributional information, particularly the occurrence of nouns that are not members of a particular class in prototypical contexts of that particular class, which provides misleading information to classifiers. The impact of this misleading information is made apparent by the amount of FP observed in classification results. In contrast, silence has to do with the well known problem of data sparsity, which can be caused by the particular distribution of lexical, and thus strict, though informative, contexts used in cue-based classification tasks, which are often rare in any corpus of any size due to their specificity.

In our experiment, we did identify some cases of nouns correctly ruled out as members of a class when using only marked cues, which were incorrectly classified as class members after the inclusion of unmarked contexts. The slight increase of FP in our results (see Table 7) shows our method does introduce some extra noise into the classifier, although, in the overall results, this is compensated by the larger amount of nouns that were correctly classified after the inclusion of unmarked distributional information (see Tables 4 and 5).

Analyzing the additional FP observed, we identify two different cases: (i) nouns correctly classified using only marked contexts as not belonging to a class based on a borderline probability, which were

incorrectly classified as members of that class when unmarked contexts were also considered, again based on a borderline probability; and (ii) nouns correctly classified as not belonging to a class as they hardly or never occurred in class-marked contexts, but whose behavior in unmarked contexts was similar to that of members of the class being classified, thus providing contradictory information to the classifier and resulting in incorrect classification.

The first case is illustrated by a noun like *biography*, which was predicted not to be a member of the LOC class with a borderline probability score (0.47). The inclusion of unmarked contexts provided information to the classifier, which slightly changed this probability (0.56), and resulted in an incorrect classification. The noun *megalopolis* illustrates the other case. Occurring only 3 times in class-marked contexts of the INF class, this LOC noun had been correctly classified as not belonging to the INF class. Yet, its behavior in unmarked contexts showed more similarities with members of the INF class than with non-members, resulting in its incorrect classification. Illustrating two paradigmatic cases of noise in the results of the classifiers, these examples make apparent how unmarked contexts are sometimes responsible for incorrect class membership decisions, and how further improving their use in classification tasks, particularly in the case of "borderline" classification decisions, remains a promising line of research to explore in the future (see Section 7).

6.2 More robust classification decisions

Besides the reduction of the impact of silence in the results of the classifiers, with the consequent improvements in accuracy, as discussed in the previous section, we also noticed that the introduction of unmarked contexts provided additional information regarding the distribution of nouns that were classified by chance (i.e. correctly classified nouns, with a borderline probability score), resulting in more robust classification decisions. We saw this with the EVT noun *consolidation* and the LOC noun *coalfield*. Each of these nouns was correctly classified using only marked contexts, yet with borderline probability scores: 0.52 and 0.53, respectively. Upon providing information on unmarked contexts to the classifier, these nouns continued to be correctly classified but with much higher probability scores, and thus more reliable: 0.75 and 0.76, respectively.

These examples are considerably different from those discussed in Section 6.1, as these are far from being cases of silence. In fact, the EVT noun *consolidation* occurs 312 times in the corpus and 317 times in marked contexts while the LOC noun *coalfield* occurs 52 times in the corpus and 53 times in marked contexts[4]. In both cases, almost all of the occurrences in marked contexts were found to be in only one cue, which was therefore not strongly valued by the classifier, as few correlations between the evidence available could be made, hence the low probability scores observed. The inclusion of unmarked distributional information provides "bridging information", allowing for more reliable classifications, which is crucial to consider especially when the ultimate goal of improving and tuning classification systems is to employ classification results for the automatic production of language resources (see Section 1).

6.3 Classification results unevenly affected by unmarked contexts

As made apparent by the results, the contribution of unmarked contexts to the classification of different semantic classes is not always the same. For example, we observed that classes whose members demonstrated a more heterogeneous linguistic behavior, such as the ORG, LOC or EVT classes, improve more with the inclusion of unmarked distributional information than classes with a more homogeneous distributional behavior. To make our statement clearer, we claim that some nominal classes are composed of nouns that tend to occur in a wider range of contexts, thus displaying a more heterogeneous and disperse distributional behavior. This heterogeneity is made apparent by an analysis of the overall distribution of the marked cues between the members of each lexical semantic class. In contrast with heterogeneous noun classes, other classes are composed of members that display a more homogeneous collective behavior that is more easily captured by distributional approaches[5].

[4] Note that a single occurrence in corpus data can activate more than one cue considered in our experiments (for instance, in the case of a target noun that has a marked suffix and simultaneously occurs in a marked syntactic construction), hence the higher amount of occurrences in cues than overall occurrences in the corpus discussed in the examples introduced in this paragraph.

[5] Our analysis of the data showed that the dispersion of distributional behavior is independent of frequency.

Analyzing the distribution of cues between class members in Corpus B, we identified, in each class, a set of cues that occurred with the majority of nouns of the class, and which we will consider to represent the core linguistic behavior of each specific class. We also observed the amount of cues included in this set differed considerably from class to class (see Figure 1). Thus, the larger the amount of marked contexts shared by the majority of the members of a class, the more homogeneous we can claim their behavior to be.

In the specific case of the classes considered in this paper, 30.7% of the cues for the HUM class are shared by the majority of HUM nouns, while 26.6%, 13.3%, 9.5% and 9.1% of the cues for the INF, ORG, EVT and LOC classes, respectively, are shared by the majority of the nouns of these classes, as represented in Figure 1. An effect of a class collectively having a more heterogeneous linguistic behavior is that the evidence regarding each of its marks will typically be more disperse and, as a result, often not strong enough to be considered by classifiers, which explains the improvement introduced by unmarked contexts. In contrast, classes like HUM are composed of nouns that generally occur in a common set of prototypical contexts of that class. Thus, on the one hand, identifying contexts that mirror the prototypical behavior of that class is more straightforward and, on the other, class members almost always show enough occurrences in such contexts to be accurately classified.

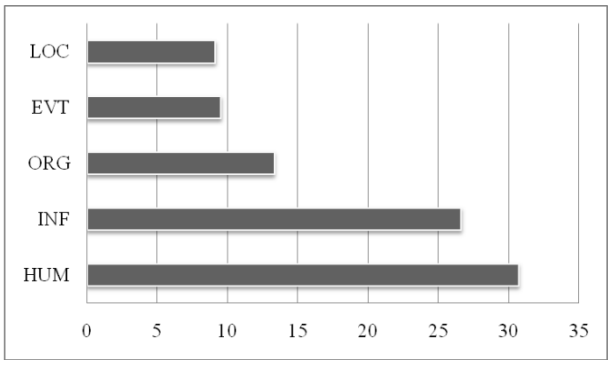

Figure 1: Percentage of cues shared by the majority of class members, per class

Additionally, there are also strong marks based on suffixes and degree of grammaticalization for the HUM class (as demonstrated in Bel *et al.* (2012)), which can be more readily captured by these more available marked contexts. For instance, on the one hand, suffixes, such as: *"-er"* and *"-or"* are indicative of many HUM type nouns (e.g. "doctor", "painter", "officer", etc.) while the preposition *"during"*, when preceding a nominal phrase, is very indicative of occurrences of EVT nouns. These examples are both instances of features that can be easily identified for inclusion in a feature vector, readily providing a large amount of class-indicative information. On the other hand, there are other types of features that although indicative, result in a much sparser vector because of their reliance of occurrence within corpus data. For instance the occurrence as the subject of an agentive verb, which is considered an indicative feature for the ORG class, does not necessarily occur readily with all members of the class, thus making marked contexts that provide a homogeneous representation of the class more difficult to capture.

In this way, the inclusion of extra contexts (e.g. unmarked contexts) are rendered ineffective when class membership decisions are already accurately made to a great extent (in our case 86.19% of the times) based on the information provided by marked contexts. This is consistent with the stability of the results reported for the HUM class in the different experiments performed, which did not demonstrate any significant changes with the inclusion of unmarked contexts.

7 Conclusions and Final Remarks

Our main goal in this paper was to evaluate how unmarked contexts can be used to improve the accuracy of nominal lexical semantic classification tasks. Departing from the hypothesis that these contexts can provide additional information to classifiers when there is not enough distinctive co-occurrence information available, the results reported demonstrate the use of unmarked contexts, which are typically discarded as non-discriminatory, can significantly improve the results of lexical semantic classification when considered along with marked contexts. Our results also show that using both types of distributional information (marked and unmarked) is crucial to reduce the sparse data problem, thus

improving classification (see increase in classification accuracy in Tables 4 and 5). Moreover, in our experiments, we apply this method to two independent corpora obtaining comparable results and thus demonstrating the robustness and transferability of our approach to any dataset.

The higher accuracy and error reduction achieved with the inclusion of unmarked contexts constitute a significant improvement with respect to the state of the art (Bel, 2010; Bel *et al.*, 2012; Romeo *et al.*, 2014), contributing particularly to the increase of accuracy and reliability of classifiers for classes that exhibit more disperse linguistic behavior. Moreover, the approach depicted here leaves room for further improvements and future work, particularly with regard to designing strategies to minimize the introduction of borderline false positives in classification.

One promising line of research to explore is the optimization of the inclusion of unmarked contexts in classification decisions. As detailed in the discussion, for the experiments depicted in this paper, we did not expect particular marked or unmarked features to be more useful than others, as we relied on the correlation of all the distributional information considered for each specific class to be indicative of class membership.

Another aspect to be further explored consists of determining the most effective amount of unmarked contexts to be provided to automatic systems. Building on the demonstration of the positive contribution of unmarked contexts in classification tasks, as indicated by the results obtained in the work depicted in this paper (see Section 5), we will start by determining the specific contribution to classification of each unmarked feature used. In this way, we would check whether there is a context, within our set, that is not contributing to the classification, in order to establish a threshold to systematically identify the information that is not relevant or whether we need to widen/relax our frequency criterion to include more unmarked contexts with the goal of elaborating a set of information to be as robust as possible, thus resulting in more accurate and more reliable classification decisions.

Finally, we believe the results obtained make a clear contribution towards the automatic production of high-quality language resources, which will benefit any NLP system that requires information on lexical semantic classes as an input.

Acknowledgements

This work was funded with the support of the SUR of the DEC of the Generalitat de Catalunya and the European Social Fund, by SKATER TIN2012-38584-C06-05, and by Fundação para a Ciência e a Tecnologia (FCT) post-doctoral fellowship SFRH/BPD/79900/2011.

References

Abbès, S. B., Zargayouna, H. and Nazarenko, A. 2011. Evaluating Semantic Classes Used for Ontology Building and Learning from Texts. In *Proceedings in the International Conference on Knowledge Engineering and Ontology Development*. Paris, France.

Agirre, E., Bengoetxea, K., Gojenola, K. and Nivre, J. 2011. Improving Dependency Parsing with Semantic Classes. In *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics, (ACL-HLT 2011).* Portland, Oregon.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation, 43*(3): 209-226.

Bel, N. 2010. Handling of Missing Values in Lexical Acquisition, In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Bel, N., Romeo, L. and Padró, M. 2012. Automatic Lexical Semantic Classification of Nouns. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey.

Bullinaria, J. A. and Levy, J. 2012. Extracting semantic representations from word co-occurrence statistics: Stoplists,stemming and svd. *Behavior Research Methods*,44:890-907.

Bybee, J. 2010. *Language, Usage and Cognition.* Cambridge University Press, Cambridge.

Cooke, N. and Gillam, L. 2008. Distributional lexical semantics for stop lists. In *Proceedings of the 2008 BCS-IRSG conference on Corpus Profiling* (IRSG'08), Anne De Roeck, Dawei Song, and Udo Kruschwitz (Eds.). British Computer Society, Swinton, UK.

Goldberg, A. E. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford University

Press: Oxford.

Harris, Z. 1954. Distributional Structure. *Word*, 10(23): 146-162.

Jakobson, R. 1971. *Selected Writings II: Word & Language*. Mouton, The Hague.

Krenn, B. and Samuelsson, C. 1997. *The Linguist's Guide to Statistics – Don't Panic*. http://nlp.stanford.edu/fsnlp/dontpanic.pdf

Landwehr, N., Hall, M. and Frank, E. 2005. Logistic Model Trees. *Machine Learning*, 95(1-2): 161-205.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4): 235-244.

Pecina, P., Toral, A., Way, A., Papavassiliou, V., Prokopidis, P. and Giagkou, M. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011)*. Leuven, Belgium.

Pustejovsky, J. 1995. *Generative Lexicon*. The MIT Press, Cambridge.

Quinlan, R. J. 1993. C4.5: *Programs for Machine Learning. Series in Machine Learning*. Morgan Kaufman: San Mateo.

Romeo, L., Lebani G. E., Bel, N. and Lenci, A. 2014 Choosing which to use? A study of distributional models for nominal lexical semantic classification. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland.: 4366-4373.

Rumshisky, A., Grinberg, V. and Pustejovsky, J. 2007. Detecting Selectional Behavior of Complex Types in Text. In *Proceedings of the 4th International Workshop on Generative Lexicon*. Paris, France.

Turney, P. D. and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141-188.

Witten, I. H. and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition, Morgan Kaufmann: San Francisco.